(19) Europäisches Patentamt
European Patent Office
Office européen des brevets

(11) EP 2 814 218 B1

(12) **EUROPEAN PATENT SPECIFICATION**

(45) Date of publication and mention of the grant of the patent:
08.07.2020 Bulletin 2020/28

(51) Int Cl.:
H04L 29/06 *(2006.01)*

(21) Application number: **14166100.9**

(22) Date of filing: **25.04.2014**

(54) **Detecting anomalies in work practice data by combining multiple domains of information**

Erkennung von Anomalien in Arbeitspraxisdaten durch Kombinieren mehrerer Domänen von Informationen

Détection d'anomalies dans des données d'exercice de travail par combinaison de multiples domaines d'informations

(84) Designated Contracting States:
AL AT BE BG CH CY CZ DE DK EE ES FI FR GB GR HR HU IE IS IT LI LT LU LV MC MK MT NL NO PL PT RO RS SE SI SK SM TR

(30) Priority: **26.04.2013 US 201313871985**

(43) Date of publication of application:
17.12.2014 Bulletin 2014/51

(73) Proprietor: **Palo Alto Research Center Incorporated**
Palo Alto, California 94304 (US)

(72) Inventors:
- **Bart, Evgeniy**
  Sunnyvale, CA California 94086 (US)
- **Liu, Juan**
  Cupertino, CA California 95014 (US)
- **Eldardiry, Hoda**
  San Carolos, CA California 94070 (US)
- **Price, Robert R.**
  Palo Alto, CA California 94306 (US)

(74) Representative: **Gill Jennings & Every LLP**
The Broadgate Tower
20 Primrose Street
London EC2A 2ES (GB)

(56) References cited:
US-A1- 2004 167 893 US-B2- 7 017 186

- **VANDANA P. JANEJA ET AL: "Multi-domain anomaly detection in spatial datasets", KNOWLEDGE AND INFORMATION SYSTEMS, vol. 36, no. 3, 18 October 2012 (2012-10-18), pages 749-788, XP055152208, ISSN: 0219-1377, DOI: 10.1007/s10115-012-0534-5**
- **Malek B Salem ET AL: "A Survey of Insider Attack Detection Research", , 25 August 2008 (2008-08-25), XP055152295, Retrieved from the Internet: URL:http://www.dtic.mil/docs/citations/ADA 519455 [retrieved on 2014-11-11]**
- **CHANDOLA V ET AL: "ANOMALY DETECTION: A SURVEY", ACM COMPUTING SURVEYS, ACM, NEW YORK, NY, US, US, 1 September 2009 (2009-09-01), pages 1-72, XP002510588, ISSN: 0360-0300 [retrieved on 2007-08-15]**

EP 2 814 218 B1

Description

BACKGROUND

Field

[0001] This disclosure is generally related to detecting malicious insider activities. More specifically, this disclosure is related to a probability model for detecting malicious activity performed by insiders within an organization.

Related Art

[0002] Malicious insiders are an increasing threat to many organizations. People authorized to access internal information may disrupt the organization's operations or leak sensitive information to outside parties. As a result, collecting and monitoring work practice data within organizations is becoming prevalent. This work practice data includes many typical activities of PC users, such as logging on/off, accessing web sites, sending and receiving emails, and accessing external devices or files. Each activity is called a "domain" or a "modality." For example, domains may include logon domain and email domain. One may detect malicious intent within an organization before the malicious activity occurs by analyzing this data and detecting anomalies. Currently, this analysis is limited to a single domain at a time. Typically, one detects anomalies or outliers separately within each domain. Some approaches combine anomaly scores in an ad hoc manner (for example, by ignoring users who are outliers in only one domain). Users who are not outliers in any of the domains may never be labeled as outliers by these analysis methods.

[0003] Thus a technical problem exists in detecting anomalies across multiple domains which, for example, may indicate potential malicious behaviour. This problem is particularly felt in large organisations where the data volume which needs to be processed in order to detect such anomalies is high.

[0004] US2004/0167893 describes a probabilistic distribution estimation apparatus which is supplied with a string of vector data as input data, and estimates, by using a stochastic model having hidden variables, a probabilistic distribution in which each data occurs by successively reading the train of vector data.

[0005] "Multi-domain anomaly detection in spatial datasets" by Vandana P. Janeja et al. describes an approach for finding a tangible anomalous window across multiple domains where window refers to the set of contiguous points in space, and since the window is multi-domain, there are several overlapping windows in the same space across domains. The approach for finding anomalous window across the domains comprises the following steps: (1) single-domain anomaly detection: discovering anomalous window in each domain; (2) association rule mining: discovering relationship between the anomalous window across domains using association rule mining; and (3) validation: validating the result using (a) Monte Carlo simulation, (b) correlation using lift and (c) ground truth evaluation.

[0006] "A Survey of Insider Attack Detection Research" by Malek Ben Salem et al. describes insider threat detection by surveying different machine learning and modeling algorithms applied to masquerade attack detection using host-based and network based audit sources.

[0007] "Anomaly Detection: A Survey" by Varun Chandola et al. describes an overview on anomaly detection.

[0008] US7017186 describes an intrusion detection system (IDS) which can include a traffic sniffer for extracting network packets from passing network traffic; a traffic parser configured to extract individual data from defined packet fields of the network packets; and, a traffic logger configured to store individual packet fields of the network packets in a database. A vector builder can be configured to generate multi-dimensional vectors from selected features of the stored packet fields. Notably, at least one self-organizing clustering module can be configured to process the multi-dimensional vectors to produce a self-organized map of clusters.

SUMMARY

[0009] Embodiments of the present invention are set out in the attached claims.

BRIEF DESCRIPTION OF THE FIGURES

[0010]

FIG. 1 presents a diagram illustrating a system for facilitating anomaly detection based on a multi-domain probability model, according to an embodiment.

FIG. 2 presents a block diagram illustrating an exemplary probability model for multi-domain clustering, according to an embodiment.

FIG. 3 presents a flowchart illustrating an exemplary process for detecting anomalies by clustering with a multi-domain probability model, according to an embodiment.

FIG. 4 presents a block diagram illustrating estimating multi-domain consistency, according to an embodiment.

FIG. 5 illustrates an exemplary system for multi-domain clustering, in accordance with one embodiment of the present invention.

[0011] In the figures, like reference numerals refer to the same figure elements.

Overview

[0012] Embodiments of the present invention solve the problem of detecting anomalous insider activity by clus-

tering users with a multi-domain probability model, and detecting anomalies by using the clustering information to compare a user's activities against activities of other users in similar roles. A domain is a source of user activity information. For example, a domain can be one of: device access, e-mail, file copy and/or access, Hypertext Transfer Protocol (HTTP) access, or logon activity. A malicious activity detection system may detect anomalous user activities by collecting user activity information from different domains, clustering the users based on the user activity information, and comparing a user's activities against other users with similar roles (also called peers). The clusters may correspond to user roles in organizational structures because different roles have different user activity patterns. A role is a person's job within an organization. For example, roles can include software engineers or lawyers.

[0013] A fundamental insight for detecting anomalous activity is that, within an organization, different user roles have different patterns for e-mailing, accessing files, etc. For example, a software engineer may access hundreds of source code and object files a day while compiling and testing a computer program. The software engineer may also send few internal e-mails daily. A lawyer may access only a few legal documents each day, and e-mail clients in other organizations often. These different roles have different e-mailing and file access patterns. The system can detect anomalous user activity by clustering users according to multiple domains, such as clustering by e-mail and file usage patterns, and then comparing a user to others with similar roles to detect anomalies.

[0014] The system described herein utilizes a multi-domain model (also called a global model) for the entire set of available domains, and finding outliers in that multi-domain model. There are two advantages to this modeling strategy. First, the system combines anomaly scores from multiple domains in a data-driven manner. Second, the strategy facilitates detection of anomalous behaviors that are not by themselves anomalous in a single domain.

[0015] The system may use the multi-domain probability model disclosed herein to cluster users. The multi-domain probability model facilitates inferring the probability distributions of domains associated with clusters and distribution of users among a number of clusters. A cluster of users is a group of users with domain characteristics that are within the probability distributions of the cluster. A domain characteristic is, for example, the number of e-mails a user sends daily. Another domain characteristic can be the number of files the user accesses daily. Each cluster includes users performing similar roles in an organization.

[0016] Each cluster is associated with a distribution for each domain. For example, each cluster can be associated with a distribution for the number of e-mails sent by users within that cluster, and a distribution for the number of files accessed by users within that cluster. To detect anomalous user activity, the system may compare a user to other users in the same role, and determine whether the user exhibits anomalous e-mail and file usage patterns.

[0017] To detect anomalous user activity, the system may also separately cluster users for each domain to associate single-domain cluster indices with each user, and then cluster the users according to the single-domain cluster indices. For example, the system may cluster users according to the average number of files accessed daily (or within any predetermined time period), and cluster users according to an average number of e-mails sent and received daily (or within any predetermined time period). The system associates each user with a single-domain cluster number for the e-mail domain, and associates each user with a single-domain cluster number for the file domain. The system then clusters the users according to the single-domain cluster numbers from the different domains, thereby generating a discrete distribution for each user. The system can then compare a user's distribution of single-domain clusters with others that have roles similar to the user to detect anomalies. Furthermore, the system can compute an anomaly score for each user for each domain, and then compute an aggregate anomaly score by weighting the separate anomaly scores for the domains.

[0018] The methods described herein utilize standard multivariate probability inference techniques to infer a joint probability distribution. A system applying the methods described herein can obtain domain data, and then use standard probability inference techniques with the disclosed probability model to determine the probability distributions of the cluster's domains, and the distribution of users among clusters. Domain data is data that describes user activity relating to a domain. The disclosed probability model is a generative model, and belongs to the general family of topic models. A topic model is a generative model for documents, and specifies a probability procedure by which one can generate documents (e.g., user and role data).

[0019] Note that one can perform a generative process associated with the disclosed probability model by sampling a cluster, and then sampling a user from the cluster. First, one samples a cluster with an associated index. The clusters correspond to users that have similar e-mail and file usage patterns. Each cluster has a set of parameters $\mathbb{Q}_t$ that determine the users that may be grouped together in the cluster, and the domain characteristics associated with these users. For example, a cluster may include software engineering types of users. Another cluster may include lawyer types. After sampling the cluster, one can sample a user from the parameters associated with the cluster. One can sample the number of files accessed or the number of e-mails sent.

[0020] The sections that follow discuss the probability model for multi-domain clustering, a technique for single-domain clustering with multi-domain analysis, a clustering process, and an example implementation of the malicious activity detection system. Note that although this

disclosure provides examples using e-mail and file access domains, embodiments of the present invention are not limited to analyzing any particular domains.

System Architecture

[0021] FIG. 1 presents a diagram illustrating a system for facilitating anomaly detection based on a multi-domain probability model, according to an embodiment. In FIG. 1, a server 102 receives domain data over a network 104. The domain data describes user activities such as e-mails sent and received and files accessed. For example, the domain data may describe activities performed by a user 106 who has a role of software engineer. User 106 may access 1000 files a day that include object files and source code, and send three e-mails daily on average. As another example, the domain data may describe activities performed by a user 108 who has a role of attorney. User 108 may access four files a day that include legal documents and send 20 e-mails daily.

[0022] Various computers and/or other electronic equipment may collect data describing user activities such as the number of e-mails that a user sends and receives within a predetermined time period, the number of files accessed within the predetermined time period, the topical content of e-mails, the number of logons within the predetermined time period, the number and type of websites visited within the predetermined time period, number of HTTP accesses within the predetermined time period, addresses for HTTP access, number of logons within the predetermined time period, and the number of websites visited within the predetermined time period.

[0023] Domain data may also include the types of files that a user accesses. For example, a software engineer may access source code and object files, but a lawyer may access word processing documents. The system may analyze the topical distributions of documents, including e-mail. The system may also determine the statistics for collected data, such as computing the averages for all of the data collected. For example, the system may compute the average number of machines accessed daily for logon.

[0024] After receiving the event data, server 102 may cluster the user activity data. This involves determining probability distributions for domains associated with clusters, and also determining distribution of users among clusters.

[0025] After the system determines the distributions, the system may utilize the model to detect anomalous user behavior. The system and/or a human operator may compare a user with that of his peers to determine whether the user behaves similarly within the domains. Users are peers if they share a job role or position. If a user does not behave similarly to other peers within a domain, then that user behavior can be labeled as anomalous. If the user activity is unpredictable, the system may label the user as anomalous. For example, if a user sends or reads less than the typical number of e-mails that others in the same role would, the system may flag the user as anomalous.

Exemplary Probability Model

[0026] FIG. 2 presents a block diagram illustrating an exemplary probability model for multi-domain clustering, according to an embodiment. The probability model depicted in FIG. 2 illustrates dependency structures between different variables of clusters. The arrows denote the dependency structure of the probability model. Note that one can remove or add dependencies and nodes to adapt the model to suit different sets of domains.

[0027] In FIG. 2, each node (e.g., circle) corresponds to a variable in the probability model. Nodes 202a, 202c are variables representing the distribution for domains associated with clusters. Node 202b represents the role of user u, and corresponds to a cluster associated with user u. The system receives domain data represented by nodes 204a, 204b. Then, based on the domain data received, the system determines the probability distributions of the latent variables represented by nodes 202a-202c. Similar to the dependencies, one can change or remove the nodes to adapt to different domains. R represents a number of total number of roles, and 1/R represents one of the job roles which is associated with the user. "Mult" indicates that this is a multi-domain probability models. Note that users can be associated with more than one role, since users can have multiple job titles and responsibilities in organizations.

[0028] In FIG. 2, α^1 and α^2 are hyperparameters. In Bayesian statistics, a hyperparameter is a parameter of a prior distribution. A prior distribution is a probability distribution that expresses one's uncertainty about a parameter or latent variable of a distribution. The prior distribution is typically a subjective assessment of an expert. In this diagram, α^1 represents a parameter of a prior distribution for π_r^1 associated with a first domain, and α^2 represents a parameter of a prior distribution for π_r^2 associated with a second domain. One can estimate the values of α^1 and α^2. For nodes 204a, 204b, i_u^1 and i_u^2 represent the observed data for a first domain and a second domain, respectively. One may also provide a hyperparameter (not pictured) and estimate a prior distribution for r_u, which represents the distribution of users among clusters that correspond to roles.

Exemplary Process

[0029] FIG. 3 presents a flowchart illustrating an exemplary process for detecting anomalies by clustering with a multi-domain probability model, according to an embodiment. During operation, the system initially collects domain data, such as a number of e-mails sent by

users and a number of files accessed by the users (operation 302). The system may itself collect the user activity data or obtain the user activity data from computers with log records or from any machine or person that monitors and collect data on user activities. A computer operator may input the event data or the machines may automatically collect such user activity data. Next, the system obtains estimated α^1 and α^2 values as parameters for prior distributions of π_r^1 and π_r^2, respectively (operation 304).

[0030] The system may obtain both estimated values through input from a human operator. The system may also obtain estimated values from previously stored data or by generating the estimated values. The system then determines cluster probability distributions for π_r^1 and π_r^2 (operation 306). The system outputs a cluster index for each user and the probability distributions for each domain. The cluster index is associated with a role (e.g., lawyer, software engineer) for the user. The probability distributions for the domains characterize the role associated with each cluster. The system may also compute a single probability estimate for each user's data by comparing a user's data with the distributions and expected values of distributions for the clusters the user is associated with. The system may then compare users with peers to detect anomalies (operation 308).

[0031] The system may apply one of the standard techniques to determine the probability distributions of domains in clusters based on the domain data. These techniques include Gibbs sampling or variational inference. Gibbs sampling is a standard method for probability inference. Gibbs sampling is a Markov chain Monte Carlo (MCMC) algorithm for obtaining a sequence of observations from a multivariate probability distribution (e.g. from the joint probability distribution of two or more variables). The system may utilize this sequence to approximate the joint distributions. With variational inference, the system approximates the posterior distribution over a set of unobserved variables given some data (e.g., approximating the domain distributions after observing the domain evidence).

[0032] Note that embodiments of the present invention are not limited to utilizing Gibbs sampling or variational inference, and the system may also utilize other algorithms for inferring the probability distributions.

[0033] After determining the probability distributions of the clusters, the system may gauge the accuracy of the probability model. The system can generate fictitious users from the inferred probabilities, and compare the generated users to actual users to determine whether the model is accurate.

Estimating Multi-Domain Consistency

[0034] FIG. 4 presents a block diagram illustrating estimating multi-domain consistency, according to an embodiment. In some embodiments, instead of directly clustering domain data, the system may perform clustering in two stages. In a first stage, the system pre-clusters domain data for individual domains to generate single-domain clusters. In a second stage, the system clusters the single-domain clusters so that users are associated with multi-domain clusters corresponding to roles. The system may then compare a user with his peers to determine whether the user's distribution of single-domain clusters is consistent with that of his peers. If the distribution of single-domain clusters is not consistent with the peers' distribution of single-domain clusters, the system may label the user's activities as anomalous.

[0035] FIG. 4 depicts two sets of single-domain clusters corresponding to two domains. The system may initially generate single-domain clusters from domain data using standard clustering techniques. For example, the system may apply k-means clustering to cluster the domain data. Note that k-means clustering is a technique for cluster analysis which partitions a number of observations into k clusters, where each observation belongs to the cluster with the nearest mean.

[0036] The system may label e-mail domain clusters 402a, 402b, and 402c with cluster indices 1, 2, and 3, respectively. The system may also label file domain clusters 404a, 404b, and 404c with cluster indices 1, 2, and 3, respectively. The system then associates each user with a cluster index in each domain. The system clusters the user's single-domain cluster indices with a multi-domain probability graph, such as probability graph 200 from FIG. 2. For this embodiment, nodes 204a, 204b represent the individual single-domain cluster indices. Nodes 202a, 202c represent the distribution of single-domain cluster indices across the different domains for each user.

[0037] The distributions π_r^1 and π_r^2 for nodes 202a, 202c are discrete distributions, since the values in the distributions indicate indices for the single-domain clusters that the users are associated with. Note that there may be tens or hundreds of domains in some implementations, and a user may have a single domain cluster index value associated with each of the domains. α^1 and α^2 represent parameters for prior distributions of π_r^1 and π_r^2, respectively. The system may apply one of the standard techniques to determine the probability distributions. The system may then analyze the distributions to detect anomalies. Variations of how a system may analyze distributions to detect anomalies are discussed further below.

[0038] The system may determine whether user activity is anomalous by attempting to predict a user's cluster index for one domain from the user's cluster indices for other domains. In one embodiment, the system may generate a cluster vector c_u, where c_{ui} is a maximum a pos-

teriori probability (MAP) cluster index for a user u in domain i. In Bayesian statistics, a MAP estimate is the mode of the posterior distribution. In the first stage, the system may determine MAP cluster indices for each user for each domain, and add the cluster indices to vector c_u.

[0039] For a user u, domain i is consistent if the domain's cluster index c_{ui} is predictable from other cluster indices $\{c_{uj}\}_{j \neq i}$. In the simplest case, the system may use cluster indices of other users $w \neq u$ to learn a mapping from $\{c_{wj}\}_{j \neq i}$ to c_{wi}, and then determine whether this mapping generalizes to user u. With this technique, the system can determine whether domain i is consistent with the other domains for this user. If not, the system may label the user as an outlier. The system may determine an anomaly score based on overall prediction accuracy for domain i for other users. If the domain is difficult to predict in general, then the system should not penalize the incorrect predictions as severely. In contrast, for very predictable domains, the system may determine any incorrect predictions to be anomalous.

[0040] In an implementation, the system may utilize a leave-1-out technique to identify anomalous user activity. The system analyzes a specific user by fixing the domain values of all domains except for one. The basic principle is that normal individuals should be predictable. The system attempts to predict a cluster number of that domain. The system may identify the user activity as anomalous if the prediction is incorrect. For example, the system may set the domain values (e.g., cluster numbers) for a user such that logon = 1, device = 2, file = 3, and e-mail = 1. The system then attempts to predict a cluster number for the HTTP domain. If the prediction is incorrect, the system may label the user activity as anomalous.

[0041] In some embodiments, the system may compute an anomaly score for each domain, and then combine the anomaly scores. The system may combine anomaly scores across domains automatically, in a data-driven manner. As a result, the system may filter out discrepancies that are common in the data. For example, if (due to the volume of data), it is common for a user to be an outlier in at least one domain, then the system will not flag a user as anomalous overall if the user is anomalous in only one domain.

[0042] The system may compute anomaly scores for each domain and combine the anomaly scores by weighting the individual domains. The anomaly score for a domain d and user i is

$$a(d,i) = m(d,i) * \log \frac{N}{\sum_j m(d,j)},$$

where N is the total number of users and j is each user j from j=1 to N. The system may adjust the prediction miss value *m(d,i)* for each domain d to reflect the weighted value of the domain. The system may then compute an aggregate anomaly score *s(i)* for user i as $s(i) = \Sigma_d a(d,i)$.

Exemplary System For Multi-Domain Clustering Of Users

[0043] FIG. 5 illustrates an exemplary system for multi-domain clustering of users, in accordance with one embodiment of the present invention. In one embodiment, a computer and communication system 500 includes a processor 502, a memory 504, and a storage device 506. Storage device 506 stores a number of applications, such as applications 510 and 512. Storage device 506 also stores a multi-domain clustering system 508. During operation, one or more applications, such as multi-domain clustering system 508, are loaded from storage device 506 into memory 504 and then executed by processor 502. While executing the program, processor 502 performs the aforementioned functions. Computer and communication system 500 is coupled to an optional display 514, keyboard 516, and pointing device 518.

[0044] The data structures and code described in this detailed description are typically stored on a computer-readable storage medium, which may be any device or medium that can store code and/or data for use by a computer system. The computer-readable storage medium includes, but is not limited to, volatile memory, non-volatile memory, magnetic and optical storage devices such as disk drives, magnetic tape, CDs (compact discs), DVDs (digital versatile discs or digital video discs), or other media capable of storing computer-readable media now known or later developed.

[0045] The methods and processes described in the detailed description section can be embodied as code and/or data, which can be stored in a computer-readable storage medium as described above. When a computer system reads and executes the code and/or data stored on the computer-readable storage medium, the computer system performs the methods and processes embodied as data structures and code and stored within the computer-readable storage medium.

[0046] Furthermore, methods and processes described herein can be included in hardware modules or apparatus. These modules or apparatus may include, but are not limited to, an application-specific integrated circuit (ASIC) chip, a field-programmable gate array (FPGA), a dedicated or shared processor that executes a particular software module or a piece of code at a particular time, and/or other programmable-logic devices now known or later developed. When the hardware modules or apparatus are activated, they perform the methods and processes included within them.

[0047] The foregoing descriptions of various embodiments have been presented only for purposes of illustration and description. They are not intended to be exhaustive or to limit the present invention to the forms disclosed. Accordingly, many modifications and variations will be apparent to practitioners skilled in the art. Additionally, the above disclosure is not intended to limit the present invention.

Claims

1. A computer-executable method for multi-domain clustering, comprising:

   receiving, by a computing device, from one or more computing devices over a network (302) domain data for at least two domains associated with users, wherein a domain is a source type describing observable activities of a plurality of users, and wherein a respective user is associated with a respective user role in an organization;
   determining a hyperparameter value for a respective prior distribution of a probability distribution associated with a domain;
   estimating (304) a probability distribution for generating a multi-domain probability model;
   generating a multi-domain probability model that includes variables for two or more domains, based on the hyperparameter value and the estimated probability distribution;
   analyzing the domain data with the generated multi-domain probability model to assign a first user to a plurality of cluster indices associated with the first user's corresponding user role; and
   determining that an activity being performed by a second user is anomalous, based on a comparison between one or more of the second user's cluster indices to one or more of the first user's plurality of cluster indices.

2. The method of claim 1, wherein the domain data includes one or more of average number of e-mails, topical content of e-mail, average number of files, average number of HTTP access, addresses for HTTP access, and average number of logons.

3. The method of claim 1 or claim 2, further comprising:

   computing an anomaly score for, a respective user for each of the domains; and
   computing an aggregate anomaly score for the respective user that includes weighted anomaly scores for each of the domains.

4. The method of any of the preceding claims, further comprising :

   clustering domain data for a first domain;
   associating a first cluster index for the first domain with a particular user;
   clustering domain data for a second domain;
   associating a second cluster index for the second domain with the particular user; and
   generating a discrete distribution for the particular user that includes the first cluster index and the second cluster index.

5. The method of claim 4, further comprising comparing a user's cluster indices with other users' cluster indices to determine whether the user's activities are anomalous.

6. The method of claim 4, wherein associating the first cluster index for the first domain with the particular user comprises:

   estimating a maximum a posteriori probability cluster index for the user; and
   associating the user with the maximum a posteriori probability cluster index.

7. A non-transitory computer-readable storage medium storing instructions that when executed by a computer cause the computer to perform a method for multi-domain clustering, the method comprising:

   receiving, by a computing device, from one or more computing devices over a network (302) domain data for at least two domains associated with users, wherein a domain is a source type describing observable activities ofa plurality of users, and wherein a respective user is associated with a respective user role in an organization;
   determining a hyperparameter value for a respective prior distribution of a probability distribution associated with a domain;
   estimating (304) a probability distribution for generating a multi-domain probability model;
   generating a multi-domain probability model that includes variables for two or more domains, based on the hyperparameter value and the estimated probability distribution;
   analyzing the domain data with the generated multi-domain probability model to assign a first user to a plurality of cluster indices associated with the first user's corresponding user role; and
   determining that an activity being performed by a second user is anomalous, based on a comparison between one or more of the second user's cluster indices to one or more of the first user's plurality of cluster indices.

8. The non-transitory computer-readable storage medium of claim 7, wherein the domain data includes one or more of average number of e-mails, topical content of e-mail, average number of files, average number of HTTP access, addresses for HTTP access, and average number of logons.

9. The non-transitory computer-readable storage medium of claim 7 or claim 8, wherein the computer-readable storage medium stores additional instructions that, when executed, cause the computer to perform additional steps comprising:

computing an anomaly score for a respective user for each of the domains; and
computing an aggregate anomaly score for the respective user that includes weighted anomaly scores for each of the domains.

10. The non-transitory computer-readable storage medium of any of claims 7 to 9, wherein the computer-readable storage medium stores additional instructions that, when executed, cause the computer to perform additional steps comprising:

clustering domain data for a first domain;
associating a first cluster index for the first domain with a particular user;
clustering domain data for a second domain;
associating a second cluster index for the second domain with the particular user; and
generating a discrete distribution for the particular user that includes the first cluster index and the second cluster index.

11. The non-transitory computer-readable storage medium of any of claims 7 to 10, further comprising comparing a user's cluster indices with other users' cluster indices to determine whether the user's activities are anomalous.

12. The non-transitory computer-readable storage medium of any of claims 7 to 11, wherein associating the first cluster index for the first domain with the particular user comprises:

estimating a maximum a posteriori probability cluster index for the user; and
associating the user with the maximum a posteriori probability cluster index.

13. A computing system for multi-domain clustering, the system comprising:

one or more processors,
a computer-readable medium coupled to the one or more processors having instructions stored thereon that, when executed by the one or more processors, cause the one or more processors to perform the computer-executable method according to any of claims 1 to 6.

Patentansprüche

1. Computerausführbares Verfahren für ein Mehrdomänen-Clustering, umfassend:

Empfangen, durch ein Rechengerät von einem oder mehreren Rechengeräten über ein Netzwerk (302), von Domänendaten für wenigstens zwei Domänen, die mit Benutzern assoziiert sind, wobei eine Domäne ein Quelltyp ist, der beobachtbare Aktivitäten einer Vielzahl von Benutzern beschreibt, und wobei ein entsprechender Benutzer mit einer entsprechenden Benutzerrolle in einer Organisation assoziiert ist,
Bestimmen eines Hyperparameterwerts für eine entsprechende vorausgehende Verteilung einer mit einer Domäne assoziierten Wahrscheinlichkeitsverteilung,
Schätzen (304) einer Wahrscheinlichkeitsverteilung für das Erzeugen eines Mehrdomänen-Wahrscheinlichkeitsmodells,
Erzeugen eines Mehrdomänen-Wahrscheinlichkeitsmodells, das Variablen für zwei oder mehr Domänen enthält, basierend auf dem Hyperparameterwert und der geschätzten Wahrscheinlichkeitsverteilung,
Analysieren der Domänendaten mit dem erzeugten Mehrdomänen-Wahrscheinlichkeitsmodell für das Zuweisen eines ersten Benutzers zu einer Vielzahl von Clusterindizes, die mit der entsprechenden Benutzerrolle des ersten Benutzers assoziiert sind, und
Bestimmen, dass eine durch einen zweiten Benutzer durchgeführte Aktivität anomal ist, basierend auf einem Vergleich zwischen einem oder mehreren Clusterindizes des zweiten Benutzers mit einem oder mehreren Clusterindizes des ersten Benutzers.

2. Verfahren nach Anspruch 1, wobei die Domänendaten die durchschnittliche Anzahl von E-Mails, den Themeninhalt von E-Mails, die durchschnittliche Anzahl von Dateien, die durchschnittliche Anzahl von HTTP-Zugriffen, die Adressen für HTTP-Zugriffe und/oder die durchschnittliche Anzahl von Anmeldungen enthalten.

3. Verfahren nach Anspruch 1 oder 2, das weiterhin umfasst:

Berechnen einer Anomaliebewertung für einen entsprechenden Benutzer für jede der Domänen, und
Berechnen einer aggregierten Anomaliebewertung für den entsprechenden Benutzer, die gewichtete Anomaliebewertungen für jede der Domänen enthält.

4. Verfahren nach einem der vorstehenden Ansprüche, das weiterhin umfasst:

Clustern von Domänendaten für eine erste Domäne,
Assoziieren eines ersten Clusterindex für die erste Domäne mit einem bestimmten Benutzer,
Clustern von Domänendaten für eine zweite Do-

mäne,
Assoziieren eines zweiten Clusterindex für die zweite Domäne mit dem bestimmten Benutzer, und
Erzeugen einer diskreten Verteilung für den bestimmten Benutzer, die den ersten Clusterindex und den zweiten Clusterindex enthält.

5. Verfahren nach Anspruch 4, das weiterhin das Vergleichen von Clusterindizes eines Benutzers mit Clusterindizes von anderen Benutzern, um zu bestimmen, ob die Aktivitäten des Benutzers anomal sind, umfasst.

6. Verfahren nach Anspruch 4, wobei das Assoziieren des ersten Clusterindex für die erste Domäne mit dem bestimmten Benutzer umfasst:

Schätzen eines maximalen a-posteriori Wahrscheinlichkeits-Clusterindex für den Benutzer, und
Assoziieren des Benutzers mit dem maximalen a posteriori Wahrscheinlichkeits-Clusterindex.

7. Ein nicht-transitorisches, computerlesbares Speichermedium mit darauf gespeicherten Befehlen, die bei einer Ausführung durch einen Computer veranlassen, dass der Computer ein Verfahren für ein Mehrdomänen-Clustering durchführt, wobei das Verfahren umfasst:

Empfangen, durch ein Rechengerät von einem oder mehreren Rechengeräten über ein Netzwerk (302), von Domänendaten für wenigstens zwei Domänen, die mit Benutzern assoziiert sind, wobei eine Domäne ein Quelltyp ist, der beobachtbare Aktivitäten einer Vielzahl von Benutzern beschreibt, und wobei ein entsprechender Benutzer mit einer entsprechenden Benutzerrolle in einer Organisation assoziiert ist,
Bestimmen eines Hyperparameterwerts für eine entsprechende vorausgehende Verteilung einer mit einer Domäne assoziierten Wahrscheinlichkeitsverteilung,
Schätzen (304) einer Wahrscheinlichkeitsverteilung für das Erzeugen eines Mehrdomänen-Wahrscheinlichkeitsmodells,
Erzeugen eines Mehrdomänen-Wahrscheinlichkeitsmodells, das Variablen für zwei oder mehr Domänen enthält, basierend auf dem Hyperparameterwert und der geschätzten Wahrscheinlichkeitsverteilung,
Analysieren der Domänendaten mit dem erzeugten Mehrdomänen-Wahrscheinlichkeitsmodell für das Zuweisen eines ersten Benutzers zu einer Vielzahl von Clusterindizes, die mit der entsprechenden Benutzerrolle des ersten Benutzers assoziiert sind, und
Bestimmen, dass eine durch einen zweiten Benutzer durchgeführte Aktivität anomal ist, basierend auf einem Vergleich zwischen einem oder mehreren Clusterindizes des zweiten Benutzers mit einem oder mehreren Clusterindizes des ersten Benutzers.

8. Nicht-transitorisches, computerlesbares Speichermedium nach Anspruch 7, wobei die Domänendaten die durchschnittliche Anzahl von E-Mails, den Themeninhalt von E-Mails, die durchschnittliche Anzahl von Dateien, die durchschnittliche Anzahl von HTTP-Zugriffen, die Adressen für HTTP-Zugriffe und/oder die durchschnittliche Anzahl von Anmeldungen enthalten.

9. Nicht-transitorisches, computerlesbares Speichermedium nach Anspruch 7 oder 8, wobei das computerlesbare Speichermedium zusätzliche Befehle speichert, die bei einer Ausführung veranlassen, dass der Computer zusätzliche Schritte durchführt zum:

Berechnen einer Anomaliebewertung für einen entsprechenden Benutzer für jede der Domänen, und
Berechnen einer aggregierten Anomaliebewertung für den entsprechenden Benutzer, die gewichtete Anomaliebewertungen für jede der Domänen enthält.

10. Nicht-transitorisches, computerlesbares Speichermedium nach einem der Ansprüche 7 bis 9, wobei das computerlesbare Speichermedium zusätzliche Befehle speichert, die bei einer Ausführung veranlassen, dass der Computer zusätzliche Schritte durchführt zum:

Clustern von Domänendaten für eine erste Domäne,
Assoziieren eines ersten Clusterindex für die erste Domäne mit einem bestimmten Benutzer,
Clustern von Domänendaten für eine zweite Domäne,
Assoziieren eines zweiten Clusterindex für die zweite Domäne mit dem bestimmten Benutzer, und
Erzeugen einer diskreten Verteilung für den bestimmten Benutzer, die den ersten Clusterindex und den zweiten Clusterindex enthält.

11. Nicht-transitorisches, computerlesbares Speichermedium nach einem der Ansprüche 7 bis 10, das weiterhin das Vergleichen von Clusterindizes eines Benutzers mit Clusterindizes von anderen Benutzern, um zu bestimmen, ob die Aktivitäten des Benutzers anomal sind, umfasst.

12. Nicht-transitorisches, computerlesbares Speichermedium nach einem der Ansprüche 7 bis 11, wobei das Assoziieren des ersten Clusterindex für die erste Domäne mit dem bestimmten Benutzer umfasst:

Schätzen eines maximalen a-posteriori Wahrscheinlichkeits-Clusterindex für den Benutzer, und
Assoziieren des Benutzers mit dem maximalen a posteriori Wahrscheinlichkeits-Clusterindex.

13. Rechensystem für ein Mehrdomänen-Clustering, wobei das System umfasst:

einen oder mehrere Prozessoren,
ein computerlesbares Medium, das mit dem einen oder den mehreren Prozessoren gekoppelt ist und darauf gespeicherte Befehle aufweist, die bei einer Ausführung durch den einen oder die mehreren Prozessoren veranlassen, dass der eine oder die mehreren Prozessoren das computerausführbare Verfahren gemäß den Ansprüchen 1 bis 6 durchführen.

Revendications

1. Procédé exécutable par ordinateur pour une mise en grappe multi-domaine, comprenant :

la réception, par un dispositif informatique, en provenance d'un ou plusieurs dispositifs informatiques sur un réseau (302) de données de domaine pour au moins deux domaines associés à des utilisateurs, où un domaine est un type de source décrivant des activités observables d'une pluralité d'utilisateurs, et où un utilisateur respectif est associé à un rôle d'utilisateur respectif dans une organisation ;
la détermination d'une valeur d'hyperparamètre pour une distribution antérieure respective d'une distribution de probabilité associée à un domaine ;
l'estimation (304) d'une distribution de probabilité pour générer un modèle de probabilité multi-domaine ;
la génération d'un modèle de probabilité multi-domaine qui inclut des variables pour deux domaines ou plus, sur la base de la valeur d'hyperparamètre et de la distribution de probabilité estimée ;
l'analyse des données de domaine avec le modèle de probabilité multi-domaine généré pour assigner un premier utilisateur à une pluralité d'indices de grappe associés au rôle d'utilisateur correspondant du premier utilisateur ; et
la détermination qu'une activité qui est effectuée par un second utilisateur est anormale, sur la base d'une comparaison entre un ou plusieurs des indices de grappe du second utilisateur et un ou plusieurs de la pluralité d'indices de grappe du premier utilisateur.

2. Procédé selon la revendication 1, dans lequel les données de domaine incluent un ou plusieurs parmi un nombre moyen de courriels, un contenu d'actualité de courriel, un nombre moyen de fichiers, un nombre moyen d'accès HTTP, des adresses pour un accès HTTP, et un nombre moyen de connexions.

3. Procédé selon la revendication 1 ou la revendication 2, comprenant en outre :

le calcul d'un score d'anomalie pour un utilisateur respectif pour chacun des domaines ; et
le calcul d'un score d'anomalie agrégé pour l'utilisateur respectif qui inclut des scores d'anomalie pondérés pour chacun des domaines.

4. Procédé selon l'une quelconque des revendications précédentes, comprenant en outre :

la mise en grappe de données de domaine pour un premier domaine ;
l'association d'un premier indice de grappe pour le premier domaine à un utilisateur particulier ;
la mise en grappe de données de domaine pour un second domaine ;
l'association d'un second indice de grappe pour le second domaine à l'utilisateur particulier ; et
la génération d'une distribution discrète pour l'utilisateur particulier qui inclut le premier indice de grappe et le second indice de grappe.

5. Procédé selon la revendication 4, comprenant en outre la comparaison des indices de grappe d'un utilisateur avec les indices de grappe d'autres utilisateurs pour déterminer si les activités de l'utilisateur sont anormales.

6. Procédé selon la revendication 4, dans lequel l'association du premier indice de grappe pour le premier domaine à l'utilisateur particulier comprend :

l'estimation d'un indice de grappe de probabilité a posteriori maximum pour l'utilisateur ; et
l'association de l'utilisateur à l'indice de grappe de probabilité a posteriori maximum.

7. Support de stockage lisible par ordinateur non transitoire stockant des instructions qui, lorsqu'elles sont exécutées par un ordinateur, amènent l'ordinateur à effectuer un procédé pour une mise en grappe multi-domaine, le procédé comprenant :

la réception, par un dispositif informatique, en

provenance d'un ou plusieurs dispositifs informatiques sur un réseau (302) de données de domaine pour au moins deux domaines associés à des utilisateurs, où un domaine est un type de source décrivant des activités observables d'une pluralité d'utilisateurs, et où un utilisateur respectif est associé à un rôle d'utilisateur respectif dans une organisation ;
la détermination d'une valeur d'hyperparamètre pour une distribution antérieure respective d'une distribution de probabilité associée à un domaine ;
l'estimation (304) d'une distribution de probabilité pour générer un modèle de probabilité multi-domaine ;
la génération d'un modèle de probabilité multi-domaine qui inclut des variables pour deux domaines ou plus, sur la base de la valeur d'hyperparamètre et de la distribution de probabilité estimée ;
l'analyse des données de domaine avec le modèle de probabilité multi-domaine généré pour assigner un premier utilisateur à une pluralité d'indices de grappe associés au rôle d'utilisateur correspondant du premier utilisateur ; et
la détermination qu'une activité qui est effectuée par un second utilisateur est anormale, sur la base d'une comparaison entre un ou plusieurs des indices de grappe du second utilisateur et un ou plusieurs de la pluralité d'indices de grappe du premier utilisateur.

8. Support de stockage lisible par ordinateur non transitoire selon la revendication 7, dans lequel les données de domaine incluent un ou plusieurs parmi un nombre moyen de courriels, un contenu d'actualité de courriel, un nombre moyen de fichiers, un nombre moyen d'accès HTTP, des adresses pour un accès HTTP, et un nombre moyen de connexions.

9. Support de stockage lisible par ordinateur non transitoire selon la revendication 7 ou la revendication 8, dans lequel le support de stockage lisible par ordinateur stocke des instructions additionnelles qui, lorsqu'elles sont exécutées, amènent l'ordinateur à effectuer des étapes additionnelles comprenant :

le calcul d'un score d'anomalie pour un utilisateur respectif pour chacun des domaines ; et
le calcul d'un score d'anomalie agrégé pour l'utilisateur respectif qui inclut des scores d'anomalie pondérés pour chacun des domaines.

10. Support de stockage lisible par ordinateur non transitoire selon l'une quelconque des revendications 7 à 9, dans lequel le support de stockage lisible par ordinateur stocke des instructions additionnelles qui, lorsqu'elles sont exécutées, amènent l'ordinateur à effectuer des étapes additionnelles comprenant :

la mise en grappe de données de domaine pour un premier domaine ;
l'association d'un premier indice de grappe pour le premier domaine à un utilisateur particulier ;
la mise en grappe de données de domaine pour un second domaine ;
l'association d'un second indice de grappe pour le second domaine à l'utilisateur particulier ; et
la génération d'une distribution discrète pour l'utilisateur particulier qui inclut le premier indice de grappe et le second indice de grappe.

11. Support de stockage lisible par ordinateur non transitoire selon l'une quelconque des revendications 7 à 10, comprenant en outre la comparaison des indices de grappe d'un utilisateur avec les indices de grappe d'autres utilisateurs pour déterminer si les activités de l'utilisateur sont anormales.

12. Support de stockage lisible par ordinateur non transitoire selon l'une quelconque des revendications 7 à 11, dans lequel l'association du premier indice de grappe pour le premier domaine à l'utilisateur particulier comprend :

l'estimation d'un indice de grappe de probabilité a posteriori maximum pour l'utilisateur ; et
l'association de l'utilisateur à l'indice de grappe de probabilité a posteriori maximum.

13. Système informatique pour une mise en grappe multi-domaine, le système comprenant :

un ou plusieurs processeurs,
un support lisible par ordinateur couplé aux un ou plusieurs processeurs sur lequel sont stockées des instructions qui, lorsqu'elles sont exécutées par les un ou plusieurs processeurs, amènent les un ou plusieurs processeurs à effectuer le procédé exécutable par ordinateur selon l'une quelconque des revendications 1 à 6.

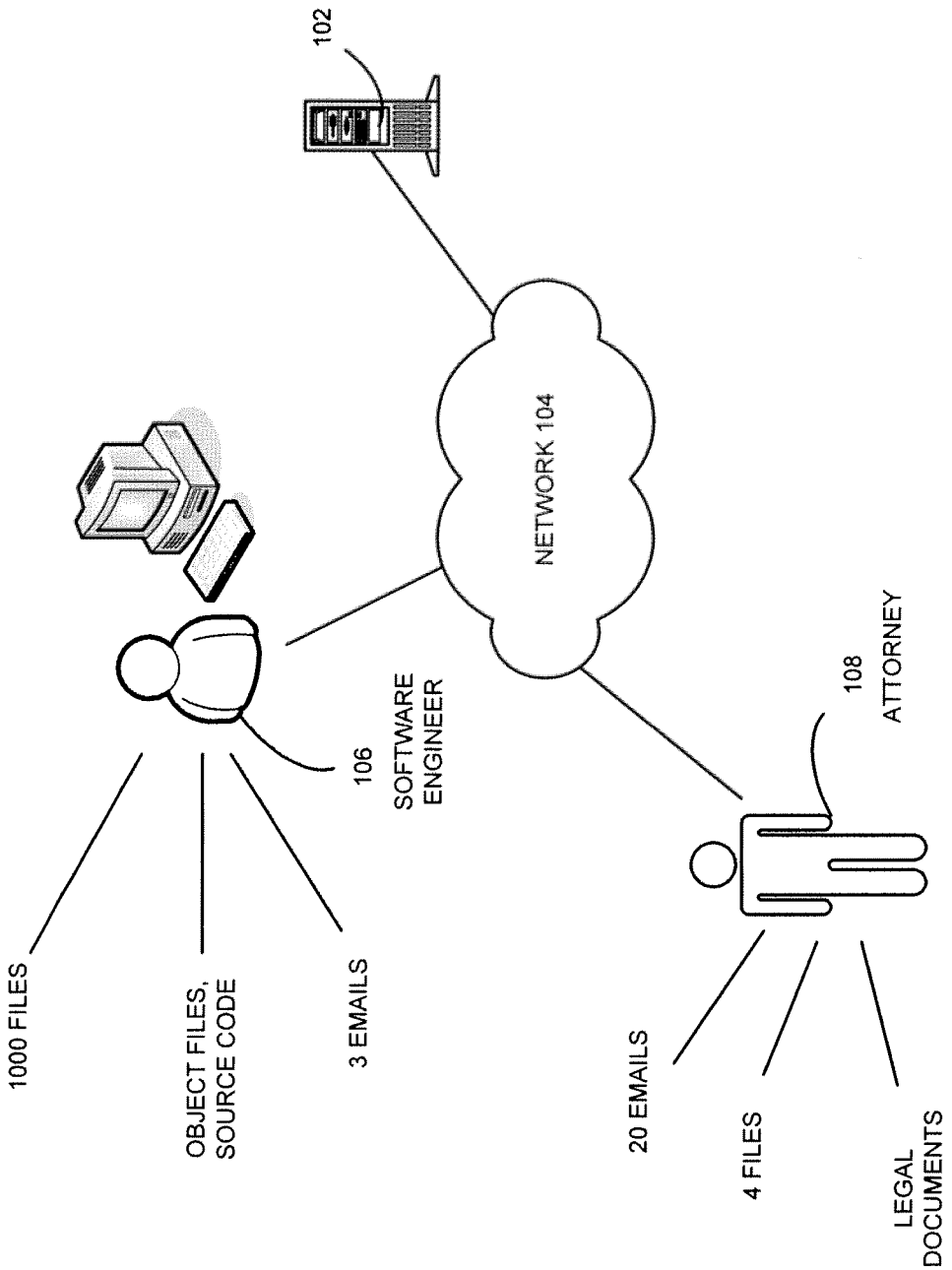
102
NETWORK 104
106
SOFTWARE ENGINEER
108
ATTORNEY
1000 FILES
OBJECT FILES, SOURCE CODE
3 EMAILS
20 EMAILS
4 FILES
LEGAL DOCUMENTS

FIG. 1

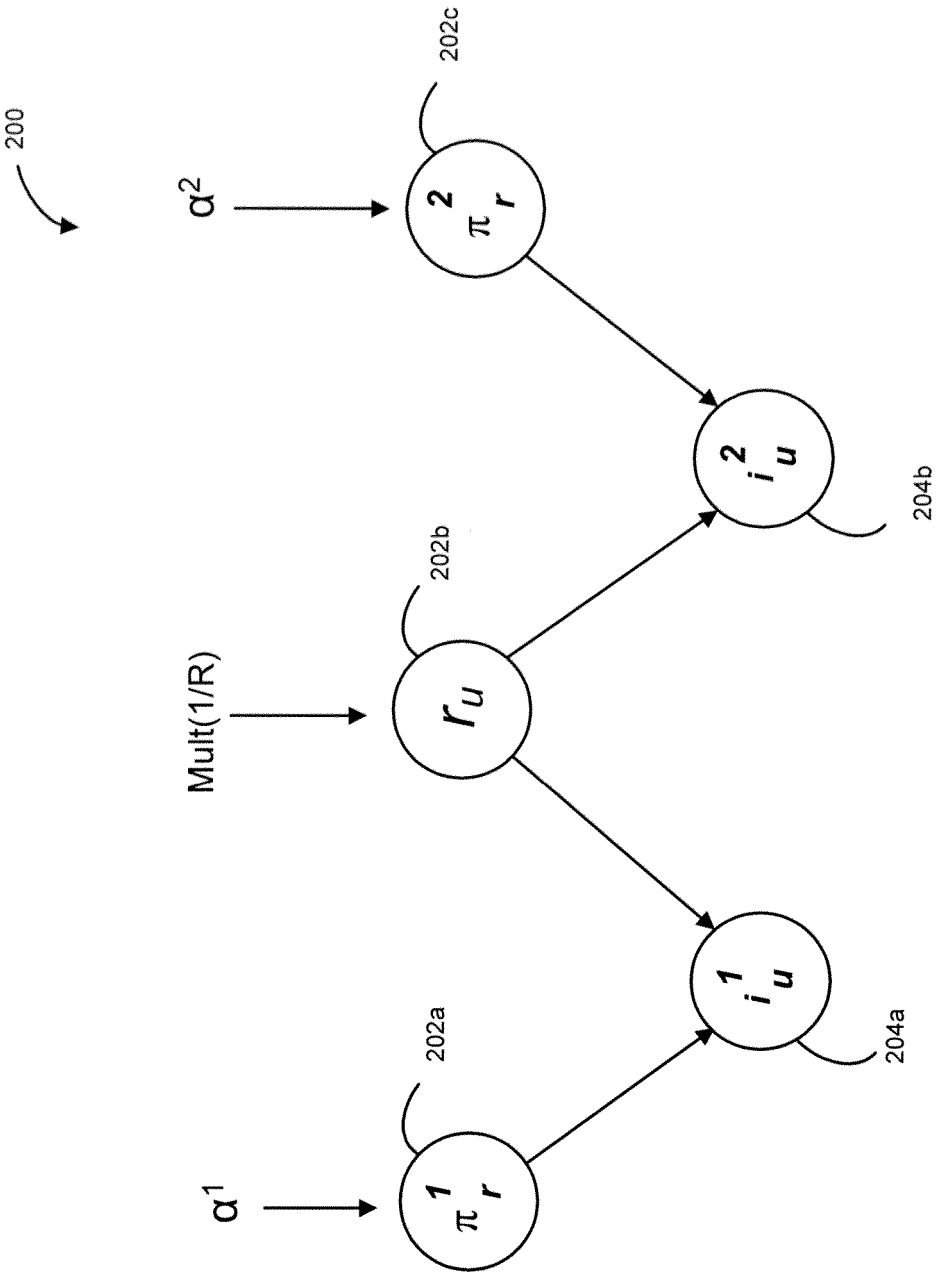

FIG. 2

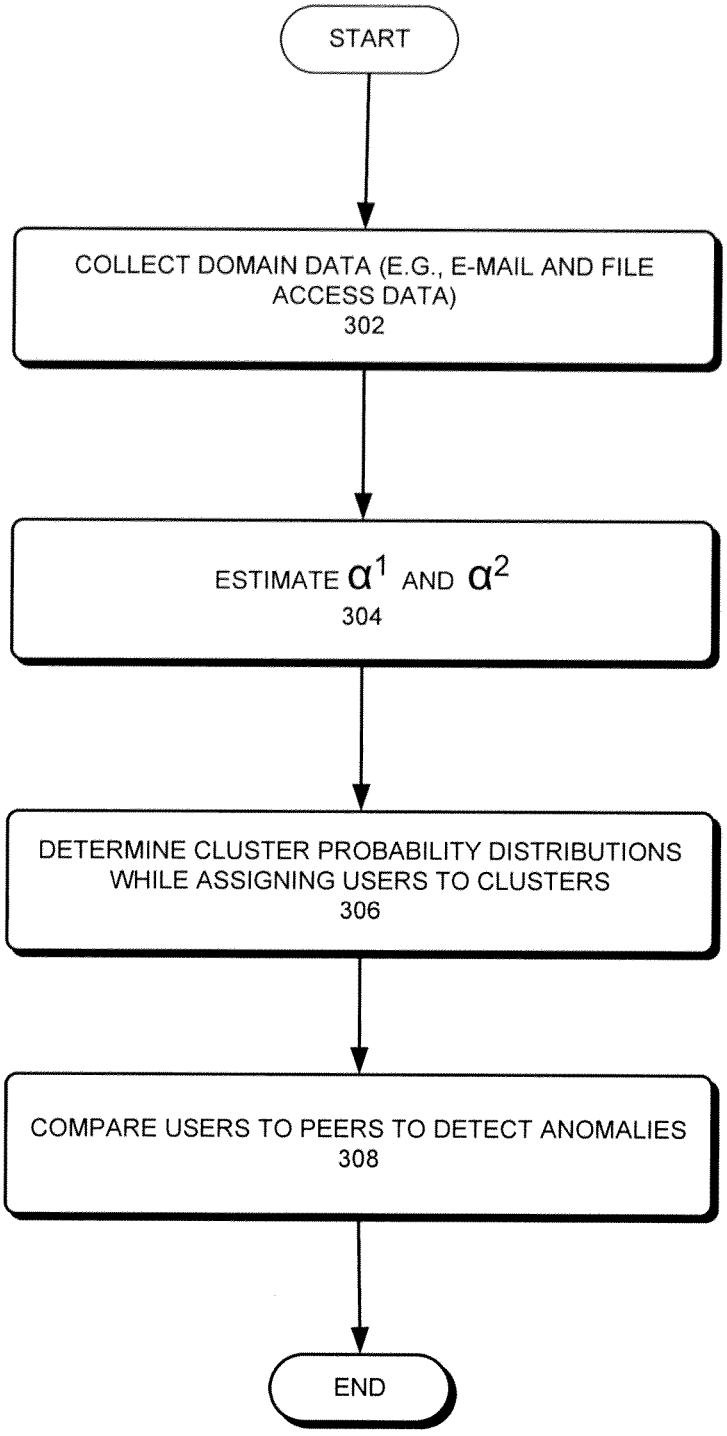
START
COLLECT DOMAIN DATA (E.G., E-MAIL AND FILE ACCESS DATA)
302
ESTIMATE α^1 AND α^2
304
DETERMINE CLUSTER PROBABILITY DISTRIBUTIONS WHILE ASSIGNING USERS TO CLUSTERS
306
COMPARE USERS TO PEERS TO DETECT ANOMALIES
308
END

FIG. 3

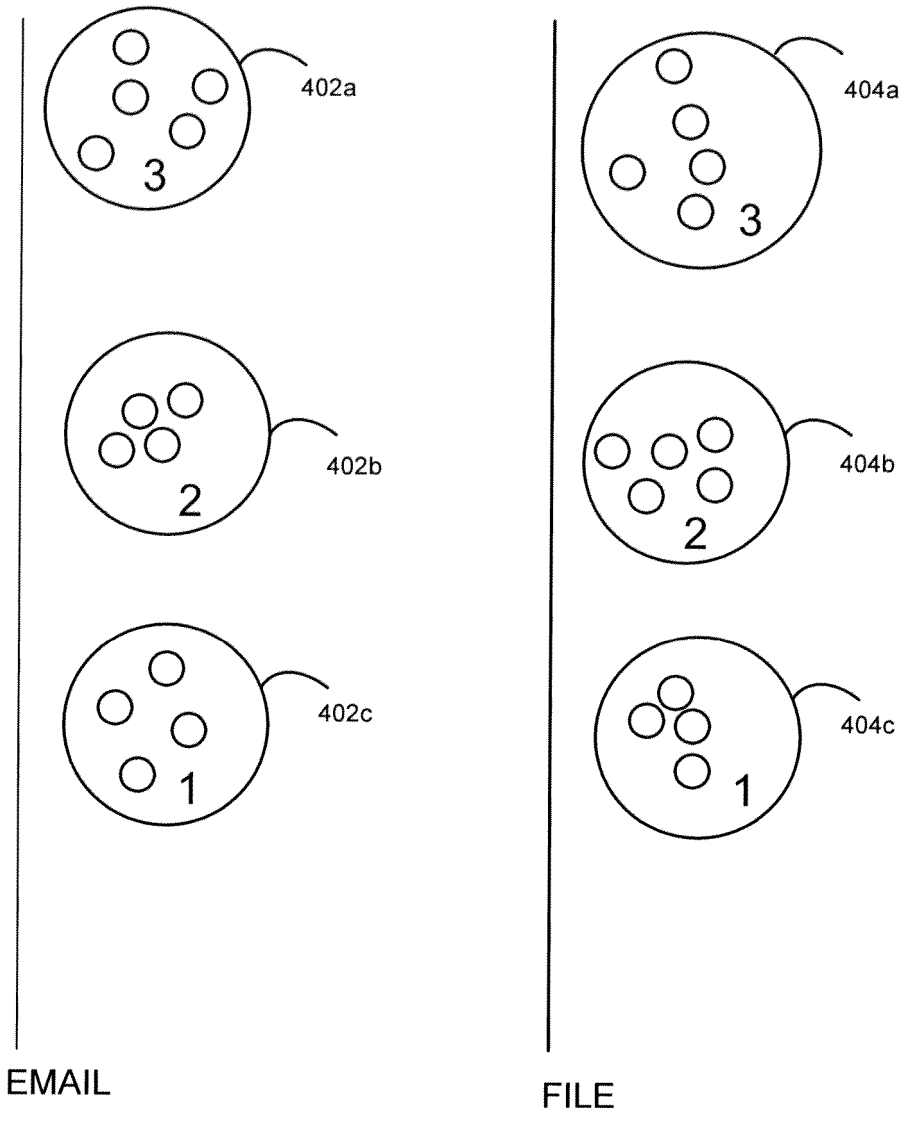

FIG. 4

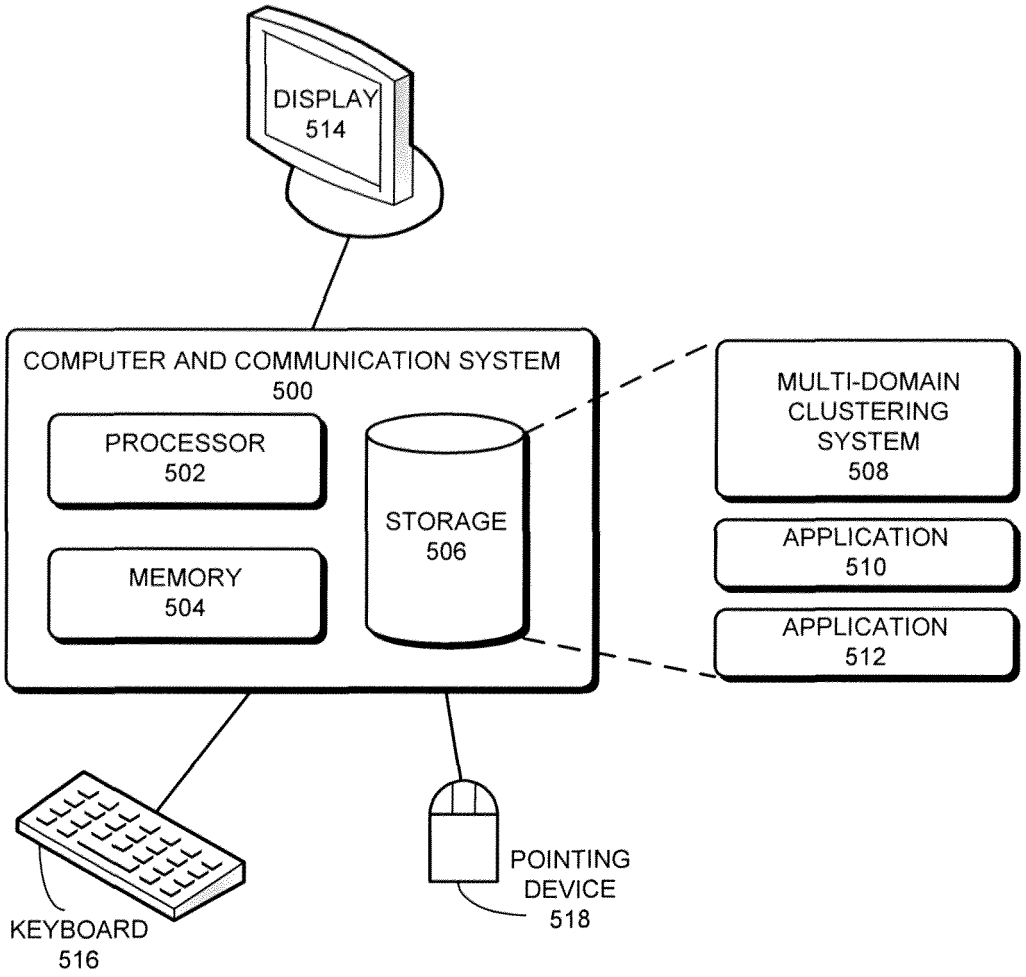
DISPLAY
514
COMPUTER AND COMMUNICATION SYSTEM
500
PROCESSOR
502
MEMORY
504
STORAGE
506
MULTI-DOMAIN
CLUSTERING
SYSTEM
508
APPLICATION
510
APPLICATION
512
KEYBOARD
516
POINTING
DEVICE
518

FIG. 5

REFERENCES CITED IN THE DESCRIPTION

This list of references cited by the applicant is for the reader's convenience only. It does not form part of the European patent document. Even though great care has been taken in compiling the references, errors or omissions cannot be excluded and the EPO disclaims all liability in this regard.

Patent documents cited in the description

- US 20040167893 A **[0004]**
- US 7017186 B **[0008]**

Non-patent literature cited in the description

- **MALEK BEN SALEM.** *A Survey of Insider Attack Detection Research* **[0006]**
- **VARUN CHANDOLA.** *Anomaly Detection: A Survey* **[0007]**